AN ALTERNATING PROJECTION ALGORITHM FOR APPROXIMATE SIMULTANEOUS DIAGONALIZATION

Riku Akema, Masao Yamagishi, and Isao Yamada

Department of Information and Communications Engineering, Tokyo Institute of Technology,
2-12-1-S3-60 Ookayama, Meguro-ku, Tokyo 152-8552, Japan
Email: {akema, myamagi, isao}@sp.ce.titech.ac.jp

ABSTRACT

In this paper, we present a novel formulation of *Approximate Simultaneous Diagonalization (ASD)* with a nonconvex feasibility problem to find a structured low rank matrix whose building blocks are the Kronecker sums of given multiple matrices. To tackle this feasibility problem, we propose an alternating projection algorithm that can generate a matrix sequence approaching monotonically to a solution. By this algorithm, simultaneously diagonalizable matrices are obtained in the neighborhood of the given matrices which are not necessarily diagonalizable simultaneously. By using further the *Diagonalize-One-then-Diagonalize-the-Other (DODO)* method, we can obtain finally a common similarity transformation which diagonalizes the simultaneously diagonalizable matrices. Numerical experiments show that, compared with a Jacobi-like method, the proposed algorithm achieves a better approximation to the desired common similarity transformation.

Index Terms— Simultaneous diagonalization, joint eigenvalue decomposition, alternating projection, nonconvex feasibility problem

1. INTRODUCTION

Exact Simultaneous Diagonalization (exact SD) [1], also called *exact Joint EigenValue Decomposition (exact JEVD)* [2, 3], is a problem to find a common similarity transformation, if exists, which diagonalizes simultaneously given multiple squared matrices.[1] For general matrices which may not have such a transformation, the goal is replaced by finding a certain common similarity transformation which diagonalizes simultaneously diagonalizable matrices (see Definition 1 in Section 2) in the neighborhood of given matrices. This type of relaxed problems is often called *Approximate Simultaneous Diagonalization (ASD)* (see, e.g., [4]). ASD has been a key for many computational strategies in data sciences including signal processing and machine learning, e.g., 2-D DOA estimation [5], joint angle-frequency estimation [6], multidimensional harmonic retrieval [7], independent component analysis [3, 8], canonical polyadic decomposition of tensors [2, 9, 10], etc.

For ASD, the so-called *Jacobi-like methods* [2, 3], which have been used extensively, are iterative algorithms for finding a minimizer in $GL(N,\mathbb{C})$ of the nonconvex function $\varphi(S) := \sum_{k=1}^{K} \mathrm{off}(S^{-1}A_kS)$, where $GL(N,\mathbb{C})(\subset \mathbb{C}^{N\times N})$ is the set of all nonsingular matrices, $A_1, A_2, \ldots, A_K \in \mathbb{C}^{N\times N}$ are given matrices, and $\mathrm{off} : X := [x_{i,j}]_{i,j=1}^{N} \mapsto \sum_{1\le i\neq j\le N} |x_{i,j}|^2$. Due to the nonconvexity of $GL(N,\mathbb{C})$ as well as φ, the Jacobi-like methods at each step do not update N^2 variables in S simultaneously, but timidly update only a single variable in certain parameterized matrices (e.g., *Givens rotation matrix* and *shear matrices*) to reach a local minimum of φ. We remark that the Jacobi-like methods do not exploit inherent algebraic properties of simultaneously diagonalizable matrices (see, e.g., Fact 1 in Section 2) to be estimated in ASD, which also suggests the possibility toward an alternative powerful strategy for ASD if we find some computational ideas to use such properties.

In this paper, we present a novel formulation of ASD with a nonconvex feasibility problem to find simultaneously diagonalizable matrices, in the neighborhood of given matrices which are not necessarily diagonalizable simultaneously, followed by an algebraic simultaneous diagonalization scheme. Unlike the Jacobi-like methods, the proposed algorithm can enjoy effectively a central property, i.e., the pairwise commutativity, of simultaneously diagonalizable matrices. To establish such a novel formulation of ASD with a nonconvex feasibility problem, we introduce a structured low rank matrix in $\mathbb{C}^{KN^2\times N^2}$ whose building blocks are the Kronecker sums of all the given matrices. We also propose an alternating projection algorithm that can generate a matrix sequence approaching monotonically to a solution of the feasibility problem. By this algorithm, simultaneously diagonalizable matrices are obtained in the neighborhood of the given matrices. By applying further the *Diagonalize-One-then-Diagonalize-the-Other (DODO)* method (found in [11]), an algebraic algorithm designed for exact SD, to the simultaneously diagonalizable matrices, we can obtain finally a common similarity transformation which is the goal of ASD. Numerical experiments show that, compared with the Jacobi-like method, the proposed algorithm achieves a better approximation to the desired common similarity transformation.

[1] This problem is different from simultaneous diagonalization by congruence (see, e.g., [1]).

Notations: Let $\mathbb{R}$ and $\mathbb{C}$ denote the set of the real numbers and the set of the complex numbers, respectively. Let $\mathrm{vec}(\cdot) : \mathbb{C}^{N\times N} \to \mathbb{C}^{N^2}$ denote vectorization by stacking columns of a matrix, and $\mathrm{vec}^{-1}(\cdot)$ its inverse. For a given $\boldsymbol{x} \in \mathbb{C}^{N^2}$, $\|\boldsymbol{x}\|$ denotes the Euclidean norm of $\boldsymbol{x}$. Let $I \in \mathbb{C}^{N\times N}$ denote the N-by-N identity matrix. For a given $X \in \mathbb{C}^{N\times N}$, $X^\top$, X^H, $\mathcal{N}(X)$, and $\|X\|_F$ denote its transpose, conjugate transpose, nullspace, and the Frobenius norm, respectively. The Kronecker product of $X = [x_{i,j}] \in \mathbb{C}^{N\times N}$ and $Y \in \mathbb{C}^{N\times N}$ is $X \otimes Y = [x_{i,j}Y]_{i,j=1}^N \in \mathbb{C}^{N^2\times N^2}$. Let $\mathcal{C}$ denote the set of all $(X_1, X_2, \ldots, X_K) \in \mathcal{X} := \mathbb{C}^{N\times N} \times \cdots \times \mathbb{C}^{N\times N}$ satisfying $X_iX_j = X_jX_i$ $(\forall i, j \in \{1, 2, \ldots, K\})$.

2. DECOMPOSITION OF APPROXIMATE SIMULTANEOUS DIAGONALIZATION INTO TWO STEPS

Definition 1. *Given multiple matrices $X_1, X_2, \ldots, X_K \in \mathbb{C}^{N\times N}$ are said to be* simultaneously diagonalizable *if there exists a common $S \in GL(N, \mathbb{C})$ such that $S^{-1}X_kS$ $(k = 1, 2, \ldots, K)$ are diagonal. We will use $\mathcal{SD}$ as the set of all $(X_1, \ldots, X_K) \in \mathcal{X}$ such that $X_1, \ldots, X_K$ are simultaneously diagonalizable.*

To establish a powerful numerical algorithm for ASD, we propose an alternative strategy which consists of the following two steps:

Step 1. for a given $\boldsymbol{A} := (A_1, A_2, \ldots, A_K) \in \mathcal{X}$, find $(\mathcal{SD} \ni)\widetilde{\boldsymbol{A}} := (\widetilde{A}_1, \widetilde{A}_2, \ldots, \widetilde{A}_K) \approx \boldsymbol{A}$;

Step 2. find $\widetilde{S} \in GL(N, \mathbb{C})$ such that $\widetilde{S}^{-1}\widetilde{A}_k\widetilde{S} (k = 1, \ldots, K)$ are diagonal.

Step 1 is introduced based on the fact that, in most data science applications of ASD, $(A_1, \ldots, A_K) \in \mathcal{X}$ are usually modeled as a slightly perturbed version of simultaneous diagonalizable matrices, say $(A_1^\star, \ldots, A_K^\star) \in \mathcal{SD}$, and that the ultimate goal of ASD is to find $S_\star \in GL(N, \mathbb{C})$ such that $S_\star^{-1}A_k^\star S_\star (k = 1, \ldots, K)$ are diagonal, where the *essential uniqueness* of $S_\star$, i.e., $S_\star$ is determined uniquely up to a permutation and a scaling of its columns, is implicitly assumed.

Despite the nonconvexity of $GL(N, \mathbb{C})$, Step 2 can be solved algebraically by the DODO method (found in [11], and also suggested by the constructive proof of Fact 1).[2]

The following characterization of $\mathcal{SD}$ is well-known.

Fact 1 (Neccesary and sufficient condition for simultaneously diagonalizable matrices [1, Theorem 1.3.21])**.** *For given diagonalizable matrices $Z_1, Z_2, \ldots, Z_K \in \mathbb{C}^{N\times N}$,*

$$\boldsymbol{Z} := (Z_1, Z_2, \ldots, Z_K) \in \mathcal{C} \Leftrightarrow \boldsymbol{Z} \in \mathcal{SD}.$$

[2]The two-step strategy was introduced in [12] but with a heuristic algorithm.

3. FORMULATION OF APPROXIMATE SIMULTANEOUS DIAGONALIZATION WITH NONCONVEX FEASIBILITY PROBLEM

3.1. Simultaneously Diagonalizability Conditions in terms of the Kronecker Sums

It is not hard to see that $X_1 \in \mathbb{C}^{N\times N}$ and $X_2 \in \mathbb{C}^{N\times N}$ commute if and only if $\mathrm{vec}(X_2) \in \mathcal{N}(I \otimes X_1 - X_1^\top \otimes I)$, where $I \otimes X_1 - X_1^\top \otimes I$ is called the *Kronecker sum* of X_1 and $-X_1^\top$. This simple fact motivates us to introduce a linear map $\Xi : \mathcal{X} \to \mathbb{C}^{KN^2\times N^2} : \boldsymbol{X} := (X_1, X_2, \ldots, X_K) \mapsto$

$$\Xi(\boldsymbol{X}) := \begin{bmatrix} I \otimes X_1 - X_1^\top \otimes I \\ I \otimes X_2 - X_2^\top \otimes I \\ \vdots \\ I \otimes X_K - X_K^\top \otimes I \end{bmatrix} \in \mathbb{C}^{KN^2\times N^2}. \quad (1)$$

Moreover, for a given $\widehat{X} \in \Xi(\mathcal{X}) := \{\Xi(\boldsymbol{Y}) \in \mathbb{C}^{KN^2\times N^2} \mid \boldsymbol{Y} \in \mathcal{X}\}$, we introduce an affine subspace $\Xi^{-1}(\widehat{X}) := \{\boldsymbol{Y} \in \mathcal{X} \mid \Xi(\boldsymbol{Y}) = \widehat{X}\} \subset \mathcal{X}$.

Lemma 1 (Characterizations of $\mathcal{C}$ and $\mathcal{SD}$ with Ξ)**.** *For a given $\boldsymbol{Z} := (Z_1, Z_2, \ldots, Z_K) \in \mathcal{X}$, we have*

(a) $\boldsymbol{Z} \in \mathcal{C} \Leftrightarrow (\forall i \in \{1, 2, \ldots, K\})\ \Xi(\boldsymbol{Z})\mathrm{vec}(Z_i) = \boldsymbol{0}$;

(b) $\boldsymbol{Z} \in \mathcal{SD} \Rightarrow$
$$\begin{cases} \mathrm{rank}(\Xi(\boldsymbol{Z})) \le N^2 - N, \\ \mathrm{rank}(\Xi(\boldsymbol{Z})) = N^2 - N \Leftrightarrow \\ \quad \textit{simultaneous diagonalization of } \boldsymbol{Z} \textit{ is ess. unique}; \end{cases}$$

(c) if at least one Z_k has N distinct eigenvalues, $\boldsymbol{Z} \in \mathcal{SD} \Leftrightarrow \mathrm{rank}(\Xi(\boldsymbol{Z})) = N^2 - N$.

(The proof of Lemma 1 is given in Appendix.)

Proposition 1 (Projection onto $\Xi^{-1}(\widehat{X})$)**.** *Let $\widehat{X} \in \Xi(\mathcal{X})$ and choose $\boldsymbol{X}^\diamond := (X_1^\diamond, \ldots, X_K^\diamond) \in \Xi^{-1}(\widehat{X})$ arbitrarily. Then,*

(a) $\Xi^{-1}(\widehat{X}) = \boldsymbol{X}^\diamond + \{(\alpha_1 I, \ldots, \alpha_K I) \mid \alpha_1, \ldots \alpha_K \in \mathbb{C}\}$;

(b) the projection onto $\Xi^{-1}(\widehat{X})$, i.e., $\mathrm{P}_{\Xi^{-1}(\widehat{X})} : \mathcal{X} \to \Xi^{-1}(\widehat{X}) : \boldsymbol{Z} := (Z_1, \ldots, Z_K) \mapsto \boldsymbol{Z}^\star :=$

$$(Z_1^\star, \ldots, Z_K^\star) := \mathop{\mathrm{argmin}}_{\boldsymbol{Y}=(Y_1,\ldots,Y_K)\in\Xi^{-1}(\widehat{X})} \sum_{k=1}^K \|Z_k - Y_k\|_F^2,$$

is given by $Z_k^\star := X_k^\diamond + \frac{\mathrm{tr}(Z_k - X_k^\diamond)}{N} I$.

Theorem 1 (Formulation of ASD with nonconvex feasibility problem)**.** *Let $\Xi(\mathcal{X}) := \{\Xi(\boldsymbol{Y}) \in \mathbb{C}^{KN^2\times N^2} \mid \boldsymbol{Y} \in \mathcal{X}\}$ and $\mathfrak{L}_{N^2-N} := \{\widehat{Y} \in \mathbb{C}^{KN^2\times N^2} \mid \mathrm{rank}(\widehat{Y}) \le N^2 - N\}$. Suppose $\widehat{X} \in \Xi(\mathcal{X}) \cap \mathfrak{L}_{N^2-N}$ and $\Xi(\boldsymbol{Z}) = \widehat{X}$ for some $\boldsymbol{Z} := (Z_1, \ldots, Z_K) \in \mathcal{X}$, where Z_1 has N distinct eigenvalues and is diagonalizable by $S \in GL(N, \mathbb{C})$ as $Z_1 = S\Lambda_1 S^{-1}$. Then, $\boldsymbol{Z} \in \mathcal{SD}$ and $S^{-1}Z_kS$ $(k = 2, \ldots, K)$ are diagonal.*

(The proof of Theorem 1 is given in Appendix.)

Theorem 1 induces us to propose the following computational strategy for Step 1 in Section 2.

Proposed Computational Strategy for Step 1

Step 1a. find $(\Xi(\mathcal{X}) \cap \mathfrak{L}_{N^2-N} \ni)\widehat{A} \approx \Xi(\boldsymbol{A})$;

Step 1b. compute $\widetilde{\boldsymbol{A}} := \mathrm{P}_{\Xi^{-1}(\widehat{A})}(\boldsymbol{A})$.

Remark 1 (On the strategy for Step 1)**.**

(a) $\widetilde{\boldsymbol{A}} = (\widetilde{A}_1, \ldots, \widetilde{A}_K)$ is expected to be close to $\boldsymbol{A}$ because $\widetilde{\boldsymbol{A}}$ is the best approximation, of $\boldsymbol{A}$, in $\Xi^{-1}(\widehat{A})$ with $\widehat{A} \approx \Xi(\boldsymbol{A})$. Moreover, with high probability, $\widetilde{\boldsymbol{A}}$ is expected to have some $\widetilde{A}_k$ of N distinct eigenvalues because of the implicit model assumption of $\boldsymbol{A}$ and [9, Theorem 6.1].

(b) The proposed computational strategy exploits an algebraic property, i.e., the pairwise commutativity, of simultaneously diagonalizable matrices. Using such a property aims to achieve a denoising effect in Step 1. The effectiveness of using the pairwise commutativity condition, for denoising in ASD, was suggested in [5] but only for $K = 2$.

3.2. Approximate Simultaneous Diagonalization Algorithm by Alternating Projection

We have already shown how to compute Step 1b in Proposition 1(b). To realize Step 1a, we propose to use an alternating projection algorithm below:

$$\left.\begin{array}{l} \widehat{A}(0) := \Xi(\boldsymbol{A}) \\ \widehat{A}(t+1) := \mathrm{P}_{\Xi(\mathcal{X})} \circ \mathrm{P}_{\mathfrak{L}_{N^2-N}}(\widehat{A}(t)) \quad (t = 0, 1, \ldots) \end{array}\right\}, \quad (2)$$

where, for any $\widehat{X} \in \mathbb{C}^{KN^2 \times N^2}$,

$$\left.\begin{array}{l} \mathrm{P}_{\mathfrak{L}_{N^2-N}}(\widehat{X}) \in \underset{\widehat{Y} \in \mathfrak{L}_{N^2-N}}{\operatorname{argmin}} \|\widehat{X} - \widehat{Y}\|_F \\ \mathrm{P}_{\Xi(\mathcal{X})}(\widehat{X}) := \underset{\widehat{Y} \in \Xi(\mathcal{X})}{\operatorname{argmin}} \|\widehat{X} - \widehat{Y}\|_F \end{array}\right\}. \quad (3)$$

Proposition 2 (Monotone approximation property of alternating projection)**.** *Let $(\widehat{A}(t))_{t=0}^{\infty}$ be the sequence generated by* (2)*. Then, the sequence $(\widetilde{\boldsymbol{A}}(t))_{t=0}^{\infty}$ defined by $\widetilde{\boldsymbol{A}}(t) := \mathrm{P}_{\Xi^{-1}(\widehat{A}(t))}(\boldsymbol{A}) \in \mathcal{X}$ satisfies $\Xi(\widetilde{\boldsymbol{A}}(t)) = \widehat{A}(t) \in \Xi(\mathcal{X})$ $(t = 0, 1, \ldots)$ and*

$$\begin{aligned} &\|\Xi(\widetilde{\boldsymbol{A}}(t+1)) - \mathrm{P}_{\mathfrak{L}_{N^2-N}}(\Xi(\widetilde{\boldsymbol{A}}(t+1)))\|_F \\ &\le \|\Xi(\widetilde{\boldsymbol{A}}(t)) - \mathrm{P}_{\mathfrak{L}_{N^2-N}}(\Xi(\widetilde{\boldsymbol{A}}(t)))\|_F \quad (t = 0, 1, \ldots). \end{aligned}$$

Remark 2 (On the proposed alternating projection)**.**

Algorithm 1 (ASD of $\boldsymbol{A}$ by alternating projection)

1: Initialize $\widehat{A}(0) = \Xi(\boldsymbol{A})$ and $t = 0$;
2: **while** $\|\widehat{A}(t) - \mathrm{P}_{\mathfrak{L}_{N^2-N}}(\widehat{A}(t))\|_F > \varepsilon$ **do**
3: $\quad \widehat{A}(t+1) = \mathrm{P}_{\Xi(\mathcal{X})} \circ \mathrm{P}_{\mathfrak{L}_{N^2-N}}(\widehat{A}(t))$;
4: $\quad t = t + 1$;
5: **end while**
6: $\widetilde{\boldsymbol{A}} := \mathrm{P}_{\Xi^{-1}(\widehat{A}(t))}(\boldsymbol{A})$;
7: Apply the DODO method to $\widetilde{\boldsymbol{A}}$;

(a) The projection $\mathrm{P}_{\mathfrak{L}_{N^2-N}} : \mathbb{C}^{KN^2 \times N^2} \to \mathfrak{L}_{N^2-N}$ in (3) *can be computed with the truncated singular value decomposition (see the* Schmidt approximation theorem*, e.g., in [13]). Note that $\mathrm{P}_{\mathfrak{L}_{N^2-N}}$ in* (3) *is determined uniquely except in a very special case where the $(N^2 - N)$-th and $(N^2 - N + 1)$-st singular values of $\widehat{X}$ happen to coincide [13].*

(b) The projection $\mathrm{P}_{\Xi(\mathcal{X})} : \mathbb{C}^{KN^2 \times N^2} \to \Xi(\mathcal{X}) : \widehat{X} := [\widehat{X}_1^\top, \ldots, \widehat{X}_K^\top]^\top \mapsto \widehat{Z} := [\widehat{Z}_1^\top, \ldots, \widehat{Z}_K^\top]^\top$ $(\widehat{X}_k, \widehat{Z}_k \in \mathbb{C}^{N^2 \times N^2},\ k = 1, \ldots, K)$ in (3) *can be computed by assigning the orthogonal projection of $\widehat{X}_k$ $(k = 1, \ldots, K)$, onto $\operatorname{span}\{I \otimes E(i,j) - E(i,j)^\top \otimes I \mid (i,j) \in (\{1, \ldots, N\} \times \{1, \ldots, N\}) \setminus \{(N,N)\}\}$, to $\widehat{Z}_k$, where $E(i,j) = [e_{l,m}^{(i,j)}] \in \mathbb{C}^{N \times N}$ is given by*

$$e_{l,m}^{(i,j)} := \begin{cases} 1 & \text{if } (l,m) = (i,j); \\ 0 & \text{otherwise.} \end{cases}$$

(c) The method of alternating projections is a powerful tool to solve feasibility problems. Even for nonconvex feasibility problems, alternating projection has a guarantee to converge locally [14, 15] to a point in the intersection and has been used extensively for finding a point, near the initial guess, in the intersection, e.g., phase retrieval [16] and Structured Low Rank Approximation (SLRA) *[17,18] (The problem in Step 1a in Section 3.1 can be seen as an instance of SLRA).*

Now, the proposed overall algorithm for ASD is given in Algorithm 1.

4. NUMERICAL EXPERIMENTS

We examine the effectiveness of the proposed algorithm. Let $\boldsymbol{A} = (A_1, \ldots, A_K) \in \mathcal{X}$ be given by

$$A_k := \frac{S_\star \Lambda_k S_\star^{-1}}{\|S_\star \Lambda_k S_\star^{-1}\|_F} + \sigma \frac{\mathcal{E}_k}{\|\mathcal{E}_k\|_F} \in \mathbb{R}^{N \times N} \ (k = 1, \ldots, K),$$

where the entries of $S_\star \in GL(N, \mathbb{R})$, a diagonal Λ_k, and an additive noise $\mathcal{E}_k \in \mathbb{R}^{N \times N}$ are drawn from a standard normal distribution, and the Signal to Noise Ratio (SNR) is defined as $-20 \log_{10}(\sigma)$. We compare the proposed algorithm (Algorithm 1) and JDTM [2]. We choose $\varepsilon = 10^{-4}$

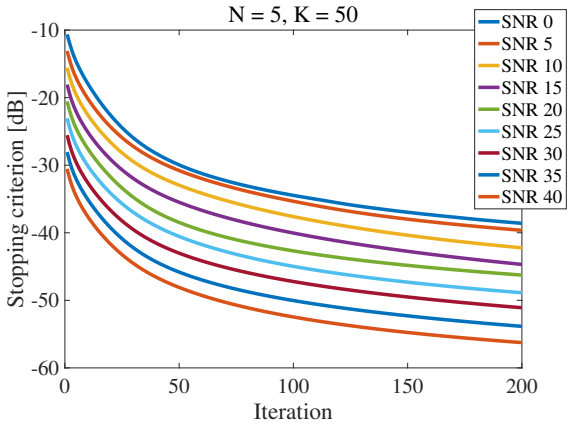


Fig. 1: Average of $(\|\widehat{A}(t) - \mathrm{P}_{\mathfrak{L}_{N^2-N}}(\widehat{A}(t))\|_F)_{t=0}^{200}$ over 100 trials.

in Algorithm 1. All the algorithms are terminated when they have reached 200 iterations. Figure 1 shows that $(\|\widehat{A}(t) - \mathrm{P}_{\mathfrak{L}_{N^2-N}}(\widehat{A}(t))\|_F)_{t=0}^{\infty}$ of Algorithm 1 decreases monotonically in all SNR settings, which is guaranteed by Proposition 2. Figure 2(a) illustrates that Algorithm 1 computes $\widetilde{\boldsymbol{A}} \in \mathcal{SD}$ closer to $\boldsymbol{A}$ than JDTM does.[3]Figure 2(b) shows that Algorithm 1 achieves significantly better approximation to $S_\star$ than JDTM in the case where (SNR)≥ 5.

5. APPENDIX

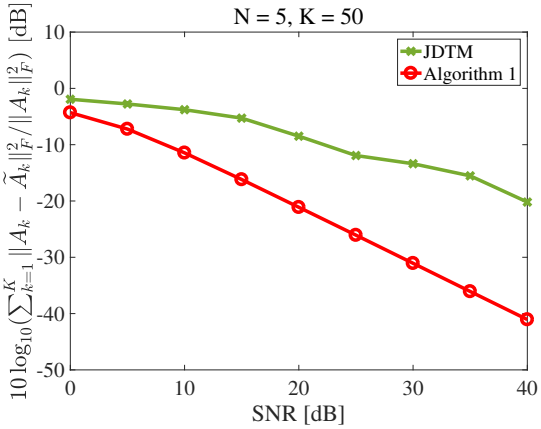
Proof of Lemma 1:

(a) This follows from the expression of the condition $Z_iZ_j - Z_jZ_i = O$ $(i, j = 1, \ldots, K)$ in vector form.

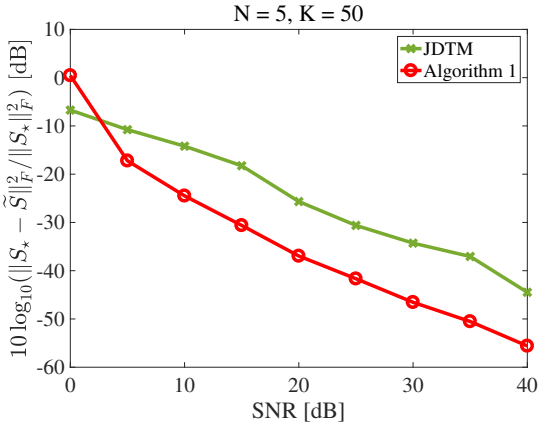
(b) Let $Z_k = S\Lambda_k S^{-1}$, where $\Lambda_k \in \mathbb{C}^{N\times N}$ is a diagonal matrix $(k = 1, 2, \ldots K)$. By using identities: $(A \otimes B)(C \otimes D) = AC \otimes BD$ $(A, B, C, D \in \mathbb{C}^{N\times N})$ [19, Lemma 4.2.10] and $(A\otimes B)^{-1} = A^{-1} \otimes B^{-1}$ $(A, B \in GL(N, \mathbb{C}))$ [19, Collorary 4.2.11], we get $I \otimes Z_k - Z_k^\top \otimes I = (S^{-\top} \otimes S)(I \otimes \Lambda_k - \Lambda_k \otimes I)(S^{-\top} \otimes S)^{-1}$. By using [19, Corollary 4.4.15], we see $\mathrm{rank}(I \otimes Z_k - Z_k^\top \otimes I) = \mathrm{rank}(I \otimes \Lambda_k - \Lambda_k \otimes I) \leq N^2 - N$. Moreover, $\Xi(\boldsymbol{Z}) = (I_K \otimes (S^{-\top} \otimes S))\Xi(\boldsymbol{\Lambda})(S^{-\top} \otimes S)^{-1}$, where $I_K \in \mathbb{C}^{K\times K}$ denotes the identity matrix and $\boldsymbol{\Lambda} := (\Lambda_1, \ldots, \Lambda_K) \in \mathcal{X}$. Now, by noting that $\Xi(\boldsymbol{\Lambda})$ has at least N zero column vectors, we deduce $\mathrm{rank}(\Xi(\boldsymbol{Z})) = \mathrm{rank}(\Xi(\boldsymbol{\Lambda})) \leq N^2 - N$. Finally, the remaining equivalence is verified by [9, Theorem 6.1].

(c) Suppose that Z_1 has N distinct eigenvalues. Therefore, there is $S \in GL(N, \mathbb{C})$ and a diagonal matrix $\Lambda_1 \in \mathbb{C}^{N\times N}$ such that $Z_1 = S\Lambda_1 S^{-1}$.

[3]In JDTM, we employ $\widetilde{A}_k := \widetilde{S}_J D_k \widetilde{S}_J^{-1}$, where $\widetilde{S}_J \in GL(N, \mathbb{C})$ is an approximation of $S_\star$ computed by JDTM and $D_k \in \mathbb{C}^{N\times N}$ is a diagonal matrix of common diagonal entries with $\widetilde{S}_J^{-1} A_k \widetilde{S}_J$.



(a) Median of normalized squared error for $\boldsymbol{A}$ over 100 trials.



(b) Median of normalized squared error for $S_\star$ over 100 trials.

Fig. 2: Comparison of JDTM [2] and the proposed algorithm (Algorithm 1) on tasks in Step 1 in Section 2 and ASD.

(Proof of "⇒") [19, Corollary 4.4.18] ensures that Z_k $(k = 2, \ldots, K)$ is a certain polynomial in Z_1. Moreover, [19, Corollary 4.4.15] ensures $\dim(\mathcal{N}(I \otimes Z_1 - Z_1^\top \otimes I)) = N$. Since $\mathcal{N}(I \otimes Z_1 - Z_1^\top \otimes I) \subseteq \mathcal{N}(I \otimes Z_k - Z_k^\top \otimes I)$ for any $k \in \{2, \ldots, K\}$, $\mathcal{N}(\Xi(\boldsymbol{Z})) = \mathcal{N}(I \otimes Z_1 - Z_1^\top \otimes I)$. Therefore, $\mathrm{rank}(\Xi(\boldsymbol{Z})) = N^2 - \dim(\mathcal{N}(\Xi(\boldsymbol{Z}))) = N^2 - N$. (Proof of "⇐") By [19, Collorary 4.4.15], $\dim(\mathcal{N}(I \otimes Z_1 - Z_1^\top \otimes I)) = N$. Therefore, $\mathcal{N}(I \otimes Z_k - Z_k^\top \otimes I) \supseteq \mathcal{N}(\Xi(\boldsymbol{Z})) = \mathcal{N}(I \otimes Z_1 - Z_1^\top \otimes I) \ni \mathrm{vec}(Z_1)$ for any $k \in \{2, \ldots, K\}$, which ensures the commutativity of Z_1 and Z_k. By using again [19, Collorary 4.4.18], we see that Z_k is a certain polynomial in Z_1. Therefore, $S^{-1}Z_kS$ $(k = 2, \ldots, K)$ are diagonal. □

Proof of Theorem 1:
From Lemma 1 (c), it is sufficient to show $\mathrm{rank}(\Xi(\boldsymbol{Z})) = N^2 - N$. Since Z_1 has N distinct eigenvalues, [19, Collorary 4.4.15] and $\Xi(\boldsymbol{Z}) \in \mathfrak{L}_{N^2-N}$ ensure $N^2 - N = \mathrm{rank}(I \otimes Z_1 - Z_1^\top \otimes I) \leq \mathrm{rank}(\Xi(\boldsymbol{Z})) \leq N^2 - N$. □

6. REFERENCES

[1] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge Univ. Press, 2nd edition, 2012.

[2] X. Luciani and L. Albera, "Canonical polyadic decomposition based on joint eigenvalue decomposition," *Chemometrics and Intelligent Laboratory Systems*, vol. 132, pp. 152–167, 2014.

[3] X. Luciani and L. Albera, "Joint eigenvalue decomposition of non-defective matrices based on the LU factorization with application to ICA," *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4594–4608, 2015.

[4] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM J. Matrix Anal. Appl.*, vol. 17, no. 1, pp. 161–164, 1996.

[5] A.-J. van der Veen, P. B. Ober, and E. F. Deprettere, "Azimuth and elevation computation in high resolution DOA estimation," *IEEE Trans. Signal Process.*, vol. 40, no. 7, pp. 1828–1832, 1992.

[6] A. N. Lemma, A.-J. van der Veen, and E. F. Deprettere, "Analysis of joint angle-frequency estimation using ESPRIT," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1264–1283, 2003.

[7] M. Haardt and J. A. Nossek, "Simultaneous Schur decomposition of several nonsymmetric matrices to achieve automatic pairing in multidimensional harmonic retrieval problems," *IEEE Trans. Signal Process.*, vol. 46, no. 1, pp. 161–169, 1998.

[8] L. Albera, A. Ferréol, P. Chevalier, and P. Comon, "ICAR: a tool for blind source separation using fourth-order statistics only," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3633–3643, 2005.

[9] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 26, no. 2, pp. 295–327, 2004.

[10] F. Roemer and M. Haardt, "A semi-algebraic framework for approximate CP decompositions via simultaneous matrix diagonalizations (SECSI)," *Signal Process.*, vol. 93, no. 9, pp. 2722–2738, 2013.

[11] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, "Numerical methods for simultaneous diagonalization," *SIAM J. Matrix Anal. Appl.*, vol. 14, no. 4, pp. 927–949, 1993.

[12] R. Akema, M. Yamagishi, and I. Yamada, "An efficient greedy algorithm for finding the nearest simultaneous diagonalizable family," in *IEEE Statistical Signal Processing Workshop (SSP)*, 2018.

[13] A. Ben-Israel and T. N. E. Greville, *Generalized inverses: theory and applications*, New York: Springer-Verlag, 2nd edition, 2003.

[14] A. S. Lewis, D. R. Luke, and J. Malick, "Local linear convergence for alternating and averaged nonconvex projections," *Found. Comput. Math.*, vol. 9, no. 4, pp. 485–513, 2009.

[15] D. Noll and A. Rondepierre, "On local convergence of the method of alternating projections," *Found. Comput. Math.*, vol. 16, no. 2, pp. 425–455, 2016.

[16] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization," *J. Opt. Soc. Am.*, vol. 19, no. 7, pp. 1334–1345, 2002.

[17] M. T. Chu, R. E. Funderlic, and R. J. Plemmons, "Structured low rank approximation," *Linear algebra and its applications*, vol. 366, pp. 157–172, 2003.

[18] L. Condat and A. Hirabayashi, "Cadzow denoising upgraded: A new projection method for the recovery of Dirac pulses from noisy linear measurements," *Sampling Theory in Signal and Image Processing*, vol. 14, no. 1, pp. 17–47, 2014.

[19] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press, 1991.